DOCUMENT RESUME

ED 400 266 TM 025 366

AUTHOR Green, Kathy E.
TITLE The Use of Person Fit Statistics in Mail Surveys.
PUB DATE Apr 96
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; *Attitude Measures; *Item Response Theory; *Mail Surveys; *Reliability; Test Construction; *Validity
IDENTIFIERS Outliers; *Person Fit Measures; *Rasch Model

ABSTRACT

Person fit statistics are generated when item response theory is used to construct measures. While person fit statistics are well grounded in theory, their utility in aggregate reporting of survey data has not been demonstrated. This study evaluated effects on reliability and validity of including and excluding misfitting person response patterns, using the Rasch model. The following mail survey data sets were used: (1) responses of 3,839 adults to a survey on the effects of the women's movement; (2) responses of 271 people to a survey on self-health care attitudes; (3) responses of 555 teachers to a survey about test use; (4) responses of 410 teachers to a survey about attitudes toward research; and (5) responses of 213 college students and graduates to a survey about dissertations. Omission of misfitting persons served to increase reliability for all data sets, but had inconsistent effects on validity coefficients. All effects were small. (Contains 3 tables and 21 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 266

The Use of Person Fit Statistics in Mail Surveys

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
KATHY GREEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Kathy E. Green

University of Denver

Paper presented at the annual meeting of the American Educational Research Association, New York, April, 1996.

TM025366

BEST COPY AVAILABLE

Abstract

Person fit statistics are generated when item response theory is used to construct measures. While person fit statistics are well grounded in theory, their utility in aggregate reporting of survey data has not been evidenced. This study evaluated the effects on reliability and validity of including and excluding misfitting person response patterns. Omission of misfitting persons served to increase reliability and had inconsistent effects on validity coefficients. All effects were small.

Person fit statistics, generated when item response theory is used to construct measures, hold promise for both individual response diagnosis and data quality evaluation (Bracey & Rudner, 1992; Harnisch, 1983). Fit statistics quantify the plausibility of a person's responses to a set of items. Responses that are implausible can be examined to generate hypotheses about an individual's knowledge, motivation, attentiveness, and sincerity. Persons whose responses are implausible may be targeted for alternative methods of data collection (e.g., interview) or their responses may simply be omitted from the sample. As Thurstone and Chave (1929) said, "The labor of tabluating the data is considerable, and we are justified in eliminating those subjects who have not responded with sufficient care or interest" (p. 32). This paper has a two-fold purpose; first, to assess the effects of removing misfitting responses and respondents on reliability and validity estimates, and, second, to identify regularities in person fit that may be diagnostically useful. Five exemplar data sets were used that varied by content and number of persons and items sampled.

Fit statistics for both persons and items are produced when an explicit measurement model underlies scaling of items and persons. When the measurement model is explicit, expectations are generated from the model that can be compared to observed responses. The discrepancy between the modeled expectations and observed values forms the basis for fit statistics. Determining the fit of the data to the modeled expectations is conceptually similar to the process used in log-linear or logit analyses, or any analysis that produces expected values for data cells. Traditional test theory yields no fit statistics since no explicit statement about individual responses to

single items is provided. Item response theory yields fit statistics since the model states an expectation for the result of each person-item encounter.

The item response theory employed in this study was the Rasch model (Rasch, 1960). The Rasch model relates a person's amount of a trait, attitude, or ability to the probability of his/her response to an item via the following mathematical model:

$$\ln[P_{nix}/P_{nix-1}] = B_n - D_i - F_x$$

where P_{nix} is the probability that person n responds to item i in category x; B_n is the interval measure of person n's attitude, D_i is the interval calibration of item i's resistance, F_x is the interval calibration of moving up one category from $x-1$ to x. While B, D, and F can be any positive or negative number, a probability must fall between 0.0 and 1.0. To deal with this concern, $B-D-F$ is introduced as the exponent of the natural logarithm base e, and a ratio is formed with e^{B-D-F} as the numerator and $1+e^{B-D-F}$ as the denominator. This yields a probability between 0 and 1. The Rasch model provides item difficulty/position and person ability/attitude estimates in logits (log-odds units) that are relatively invariant over different samples. If a person with a strongly favorable attitude answers an item that is easy to agree with, the difference between attitude and item position is large and positive, and the probability of a strongly favorable response is high. If a person with an unfavorable attitude answers a hard-to-agree with item, the difference between attitude and item position is negative, and the likelihood of a favorable response is low.

How well the data fit the model can be evaluated by subtracting expected from observed responses and squaring the result. These approximately mean square distributed fit statistics are converted to approximate t's for ease of

interpretation (Wright & Masters, 1982). Once person and item logits are estimated, the discrepancy between expected and observed values can be calculated for every person-item entry in a data matrix. Discrepancies are typically summed over persons and items to yield person and item fit values.

Fit indices tell us whether responses are as expected or are suspicious. Fit indices available at the item and person level indicate whether an item fits well with the measure or whether a person's responses are so unusual that we question his sincerity. Fit values allow identification of ill-functioning items, suspicious persons, and surprising item-person combinations, and also of responses that fit too well. Responses that fit too well may suggest socially desirable responding.

Two fit statistics are produced for each person in an analysis, infit and outfit. Infit values, or weighted total person fit, is sensitive to unexpected patterns close to the person's logit position. Outfit values, or unweighted total person fit, are sensitive to responses that might be viewed as outliers.

While person fit statistics are well grounded in theory, practical applications reported in the literature are few. Harnisch and Linn (1981), Tatsuoka and Tatsuoka (1982), and Frary (1982) suggest that fit statistics can be used to identify individuals with unusual instructional histories and thus to locate types of test bias. Persons with aberrant response patterns may be misinterpreting items or may view the construct in an unusual manner, thus invalidating their score as an indicator of the construct. Wright (1977) described possible reasons for unusual response patterns on achievement tests. Adapting these descriptions for attitude measures gives us "sleepers" who get bored and are inconsistent on later items, "fumblers" who are initially

confused by the task, and "plodders" who spend too much time on each item and never get to later ones. Add to this list those who "fake" good or bad, who interpret questions in a highly creative manner, people with atypical experiences, people who fluctuate from a conservative to a liberal interpretation within an item set, those who wish to make a point by an extreme response to some item(s), and those unmotivated people who complete the task but who are so disinterested they respond almost at random. Person fit is useful in understanding an individual's score; in particular, it permits identification of invalid scores. But, is person fit crucial when our concern is with aggregate reporting of results rather than individual diagnosis? Harnisch and Linn (1981) argue that it is; at least for achievement tests. They found differential fit to be associated with instructional differences and curriculum-test divergence. Knowledge of differential fit on attitude rather than achievement measures would allow us to explore ideas about the psychological processes affecting behavior and cognition as they are associated with personal characteristics.

Doss (1981) found the accuracy of prediction of achievement to increase with removal of misfitting person responses. Schmitt and Crocker (1984) and Garcia-Quintana (1981), however, found fit statistics to be minimally related to test anxiety, gender, race, and achievement. Schmitt, Cortina, and Whitney (1993) found removal of misfitting examinees to have little consistent effect on the validities of achievement measures and supervisors' ratings; Rudner, Skagg, Bracey, and Getson (1995) concluded that person fit "has little to offer in the analysis of traditional NAEP data" (p. iv). While Kalinowski (1985) and Gable, Ludlow, and Wolf (1990) provide examples of the diagnostic use of person fit with affective rating scales, there is little demonstration

in the literature of the practical use of fit statistics for aggregate reporting rather than individual diagnosis (Reise & Flannery, 1996).

Method

Five data sets were used to generate the examples in this paper. All data were collected via mail. The first data set was a newspaper mail-in survey while the remaining four were surveys mailed to members of sampling frames.

Women's Issues. A survey querying the effects of the women's movement appeared in the lifestyle section of a local paper with a request to mail in responses. The survey contained a total of 64 items and took up most of a newspaper page. A total of 3,839 responses to the survey were received. The measure of interest contained 30 items addressing costs/benefits of the women's movement for men, women, and society.

Self-Health Care. A mail survey about self-health care attitudes was sent to a random sample of the general population in selected towns in Wyoming. The survey contained 87 items and was 11 single-sided pages in length. Responses were received from 271 people, for a 54.2% response rate. The measure of interest comprised 8 items and addressed adherence to medical advice. No follow-up mailings were used.

Teacher's Attitudes toward Tests. A statewide mail survey of teachers' attitudes toward use of tests in schools was conducted with a random sample of Wyoming teachers. Responses were received from 555 teachers, an 81% response rate. The survey contained 49 items and was 2 double-sided pages in length. The target measure was 14 items assessing attitudes toward the value of tests in instruction. Two follow-up mailings were sent to nonrespondents.

Teacher's Attitudes toward Research. A mail survey was sent to a random sample of teachers in Nebraska and Wyoming to assess teachers' attitudes toward research. The 73.5% response rate represented 41[illegible] responses to the 54-item, 4-page survey. The measure of interest contained 23 items, representing 2 facets of attitudes toward research, applicability and theoretical utility. One follow-up mailing was sent to nonrespondents.

Responsibility for Dissertation Completion. A mail survey of current students and graduates from a College of Education in Colorado asked about their experiences completing the dissertation. Responses were received from 213 people, a 91% response rate. The survey 142 questions and was 6 pages (double-sided) in length. The Responsibility Scale had 16 items. Two follow-ups were used to encourage response.

Analyses proceeded as follows. Measures were constructed using the Rasch model computer program BIGTEPS (Linacre & Wright, 1994), with misfitting items removed. Items were considered to misfit if the mean square infit exceeded 1.3 and the content did not fit well with the general tenor of the items. Items and persons were then recalibrated and persons whose responses misfit were identified. Person misfit was arbitrarily defined as a standardized mean squared residual of +2 or more standard deviations away from the expected value. Using this criterion, about 5% of the sample would be expected to misfit by chance. With these persons removed, items and persons were again recalibrated. Measure reliability and validity coefficients were then computed with and without the persons identified as misfitting using two separate calibrations. Person fit was then plotted against person logit position with the expectation of no relationship. Associations of demographic and other selected survey variables with fit were assessed using chi-square

BEST COPY AVAILABLE

statistics. In this analysis, cases were dichotomously categorized as fitting or misfitting.

Results

Table 1 provides the number of misfitting persons and reliability coefficients with and without misfitting persons for the five data sets. The proportion of misfitting person responses was only slightly above the 5% expected by chance. Reliabilities increased when misfitting persons were removed, though the effects were small. Since misfit, either in items or persons, adds noise to the measurement process, removal of person misfits reduces this noise and so should increase reliability.

Validity was assessed by correlating scores on the target measure with other measures within the survey. Table 2 provides the correlations with and without inclusion of misfitting persons. Differences in correlations with and without misfitting persons are inconsistent, and differences occur at the second and third decimal places.

Plots of fit value and logit person position presented no discernible patterns, suggesting extremity of attitude to be unrelated to fit for these data sets.

Chi-square analyses were conducted to determine whether fit was associated with categorical survey variables. Subjects were categorized as fitting or misfitting. Table 3 presents results for the women's issues data set, which was the only one for which patterns were suggested. For these data, fit was associated significantly with age and ethnicity. Younger respondents (25-34) had fewer misfitting responses than older respondents (35-44); Hispanic respondents had more misfitting responses than Caucasians. The association with relationship style is also listed in Table 3 though this

result was marginally significant at $p = .05$. Respondents describing themselves as having polarized styles, tending to be more traditional, had more misfitting responses than respondents with a more balanced style that reflected gender equity. Respondents with balanced relationships styles tended to view effects of the women's movement more positively and were more consistent in their responses. Misfit for these data seemed to be due to unusual responses to several items rather than a strongly discrepant response to a single item. This may indicate that subgroups of respondents were interpreting the measure differently. A similar pattern of scattered misfit was found for the attitudes toward research and attitudes toward testing data.

Too few misfitting persons were found in the self-care data set to make results of chi-square analyses informative. Misfitting respondents tended to have no regular doctor and no health insurance but these results were not statistically significant. Strongly unexpected responses were found in this data set to single items. Only 3 of 16 people had unexpected responses to more than one item. This suggests [redacted] possible interactions between individuals and specific items. For example, one person was strongly favorable to all self-health care items except to the item "It's usually necessary to follow doctors' instructions."

The misfitting persons responding to the doctoral dissertation survey were more often graduates than students (8:5), and were proportionally more often female than male. Misfitting responses to this scale seemed due most often to unexpected responses to single items. For example, one respondent viewed all tasks as student responsibility but felt the university had primary responsibility for scheduling the timeline for dissertation completion, while another respondent felt the university should be responsible for filing the

BEST COPY AVAILABLE

application for graduation not the student. Inconsistencies in response patterns for the remaining persons seemed to be tied to unexpected ratings over several items rather than to random responding or misunderstanding of questions.

Discussion

Reliabilities improved when persons with misfitting responses were removed from the data set. While the effects were again small, results were consistent across data sets. Removal of persons with misfitting responses from a data set had an inconsistent effect on correlations among variables. Removal of a handful of misfitting persons from a medium to large pool of cases had very little effect on numerical summaries of relationship. However, when one is interested in greater power in discerning relationships, steps such as screening data for outliers and removal of aberrant person response patterns yield increased clarity.

People whose responses violate a standard of reasonableness present us with a dilemma. We cannot assume we are assessing the same construct for them as for others in the sample. A general concern is identification of invalid responses, yet with reporting at the aggregate level, a second concern is understanding of response strategies associated with qualitative differences among subsamples. Few associations between fit and demographic variables were found in this study, possibly because only one of the five data sets had a large enough number of misfitting persons to make such analyses wowrthwhile. Fewer misfitting patterns were found for Caucasian than for minority respondents, a result partially supported by Frary (1982).

In summary, minor advantages were found with deletion of persons with misfitting response patterns for internal consistency reliability.

Associations with demographic variables were found only for the largest data set. These results suggest assessment of fit to be useful in the manner that identification of outliers is useful. Effects may be small but our analyses are clearer.

Further research may profitably address the effects of misfit for small surveys as well as further investigation of associations with demographic and other person variables for large-scale surveys. Results of small surveys may be more strongly affected by the presence of aberrant responses, and large scale surveys would provide greater power for identifying associations. Perhaps the most interesting direction for future research would involve development of a theory explaining person misfit based on task demands and person characteristics.

References

Bookstein, A., & Lindsay, A. (1989). Questionnaire ambiguity: A Rasch scaling model analysis. Library Trends, 38, 215-36.

Bracey, G., & Rudner, L. M. (1992). Person-fit statistics: High potential and many unanswered questions. Washington, DC: American Institute for Research, ERIC EDO-TM-92-5.

Doss, D. A. (1981, April). Will removing a few bad apples save the barrel? Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Frary, R. B. (1982, March). A comparison among person-fit measures. Paper presented at the annual meeting of the American Educational Research Association, New York.

Gable, R. K., Ludlow, L. H., & Wolf, M. B. (1990). The use of classical and Rasch latent trait models to enhance the validity of affective measures. Educational and Psychological Measurement, 50, 869-78.

Garcia-Quintana, R. A. (1981, March). Person-fit to the Rasch model using norm-referenced and criterion-referenced statewide data. Paper presented at the annual meeting of the Southeastern Psychological Association, Atlanta.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20(2), 191-206.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Kalinowski, A. G. (1985). Measuring clinical pain. Journal of Psychopathology and Behavioral Assessment, 7(4), 329-349.

Kluever, R. C. (1995, April). ABDs and graduates from a college of education: Responsibility, barriers, and facilitators. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Linacre, J. M., & Wright, B. D. (1991-1994). A User's Guide to BIGSTEPS. Chicago: MESA Press.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Kobenhaven: Denmarks Paedagogiske Institut. (Reprinted by University of Chicago Press, 1980).

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. Applied Measurement in Education, 9, 9-26.

Rudner, L. M., Skagg, G., Bracey, G., & Getson, P. R. (1995). Use of person-fit statistics in reporting and analyzing National Assessment of Educational Progress results. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, NCES 95-713.

Schmitt, A. P., & Crocker, L. (1984, April). The relationship between test anxiety and person fit measures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. Applied Psychological Measurement, 17, 143-150.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7(3), 215-231.

Thurstone, L. L., & Chave, E. J. (1929). The measurement of attitude. Chicago: University of Chicago Press.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Services, pp. 85-101.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-115.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Table 1. Reliabilities with and without misfitting persons.

Data Set	n	Total Group Reliability	N Misfit	% Misfit	Persons Deleted Reliability
Women's Issues	3,839	.93	279	7.3%	.94
Self-Health Care	273	.78	17	6.2%	.81
Teacher's Attitudes toward Tests	553	.70	41	7.4%	.74
Teacher's Attitudes toward Research-Scale 1	441	.79	29	6.6%	.81
-Scale 2	441	.59	38	8.2%	.63
Responsibility for Dissertation Completion	215	.76	14	6.5%	.81

Table 2. Validity coefficients with and without misfitting persons.

Data Set	Measure	Total Group Correlation	Persons Deleted Correlation
Women's Issues	Future	-.7346**	-.7321**
	Energy	.0469**	.0973**
	Mood	.0095	.0100
	Esteem	.0408*	.0434**
	Trueself	.0501*	.0560**
	Valued	.0452**	-.0729**
Self-Health Care	Importance	.1419*	.1593*
	Independence	.4515**	.4588**
	Perceived Health	-.0675	-.1080
	Environment	.1713**	.1857**
	Chance	.0227	.0111
	Personal	-.1832*	-.1728*
Teacher's Attitudes toward Tests	Purpose	.1793**	.1710**
	Use Tests	.0907*	.0996*
	Like Tests	.2636**	.2844**
	Standardized Tests Useful	.3373**	.3314**
	Inappropriate Item Use	.1057*	.1003*
	Types Tests	.1496**	.1436**
	Types Items	.0989*	.0789

Table 2. (continued)

Data Set	Measure	Total Group Correlation	Persons Deleted Correlation
Teacher's Attitudes toward Research-Scale 1	Review	.2997**	.3119**
	Conduct	.1830**	.2034**
	Present	.1458**	.1423**
	Course Qual	.2453**	.2642**
	Course Use	.3105**	.2382**
	Teach Skill	.0242	.0592
	Research Reader	.3249**	.3184**
	Research Producer	.2013**	.1846**
-Scale 2	Review	.2518**	.2487**
	Conduct	.1791**	.1732**
	Present	.1089*	.1008*
	Course Qual	.2148**	.2192**
	Course Use	.3084*	.2787**
	Teach Skill	.1050*	.1348**
	Research Reader	.2249**	.2137**
	Research Producer	.2518**	.2225**
Responsibility for Dissertation Completion	Status	.2061**	.2081**
	Emotional Support		
	-Advisor	.2481**	.2960**
	-Committee	.2358**	.2424**
	-Students	-.0009	-.0112
	Sub3	-.2099**	-.2295**
	Sub10	-.1473*	-.1779*
	Sub11	-.1446*	-.1559*
	HH	.1442*	.1645*

Table 3. Associations between survey variables and fit: Women's issues.

Variable	Chi-Square	df	p
Relationship Style-connected, automomous, balanced	5.74	2	.05
Age (7 categories)	14.46	6	.03
Ethnicity (6 categories)	20.37	5	.01